

Local Safe Havens Can Add Value to e-Research with Derived Data

Claire MacDonald (Research Manager, NHSGGC Safe Haven), Billy Sloan (Technical Specialist, NHSGGC Safe Haven)

AIMS

To consolidate data scattered across several clinical databases and create new specialised derived datasets (BMI, Ethnicity, Charlson Score, Smoking Status) which add value and impact to research.

BACKGROUND

In 2016 NICE published the first ever multimorbidity guideline (NICE, 2016). The guidelines encourages clinicians to take an approach to care which identifies and involves multimorbidity, defined as those with two or more long-term health conditions.

Safe Havens provide extracts of health data for clinical research. Often demographic and morbidity data, such as ethnicity and BMI is requested in order to describe populations and predict health patterns. Retrospective research often has difficulty with accurately describing populations due to data being incomplete.

Providing a validated morbidity score such as Charlson index with data extracts allow s cohort/population comparisons and enables research produced from the data to be more citable and relevant.

At the NHSGGC Safe Haven we undertook the task of deriving datasets for Ethnicity, BMI, Charlson Score and Smoking status by combining data that had been entered into several separate databases. Previously, it proved very difficult to provide any of this data in a reliable or accurate way due to it being very piecemeal and inconsistently collected.

METHODS

Advice was taken from clinician researchers and literature reviewed before embarking on dataset creation for BMI, Ethnicity, Charlson Score (Deyo, et al., 1992) (Quan, et al., 2005) and Smoking Status (Masters 2018) (Wood, et al., 2005). For each derived dataset, the IBM Origin SQL Server data warehouse was searched for related fields across several clinical databases (Table 1). Data were extracted into a new ‘measurements’ table. Once extraction was complete, a new Matrix table was built, starting with unique CHI numbers from the measurements found and filled with the required data using simple date matching and algorithms of varying complexity, where required, to complete the data. The team member validating the datasets was independent from the creator.

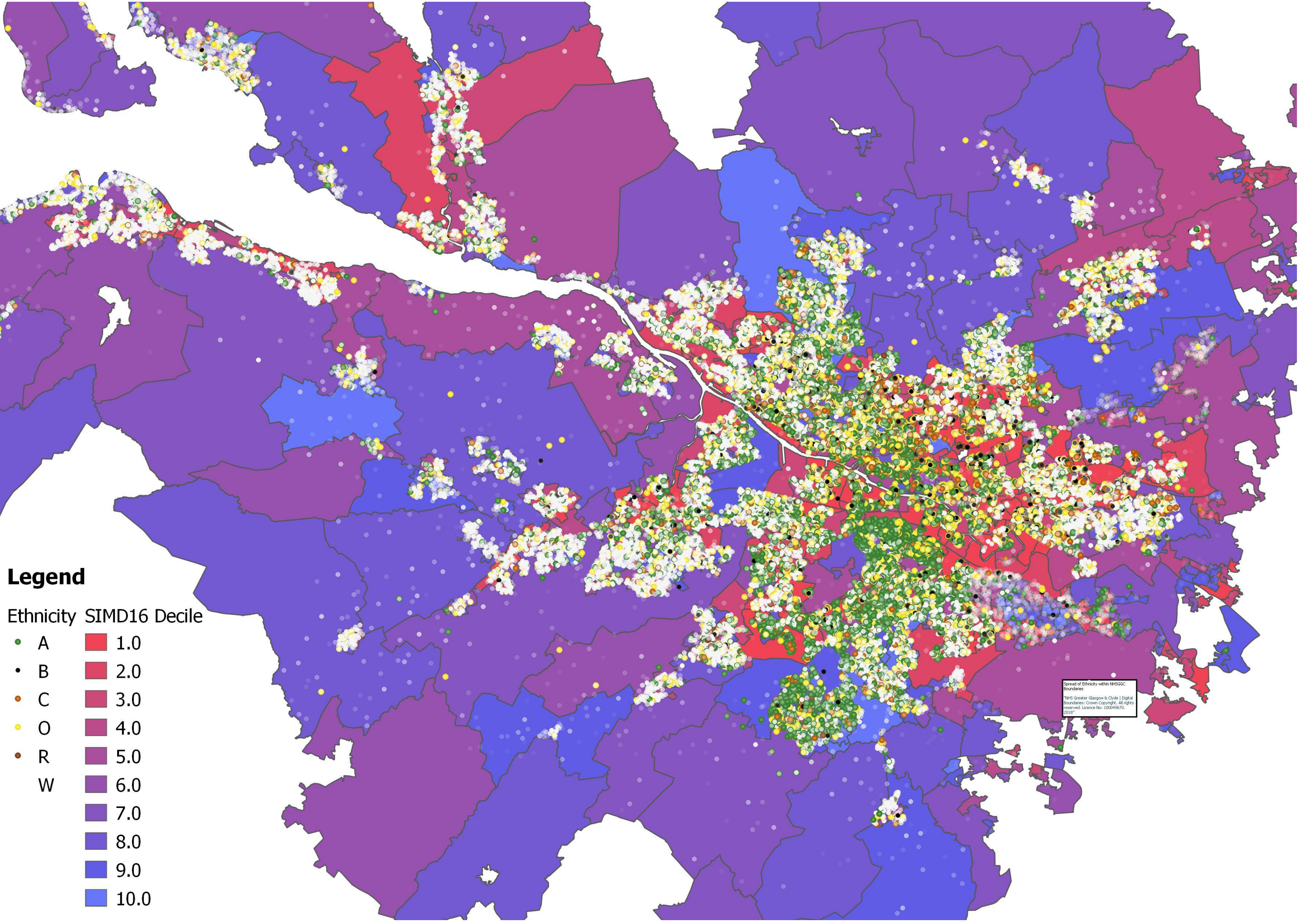


Figure 1: Map showing ethnicity spread of captured living & deceased health population across NHSGGC plotted on a background of SIMD16 Datazone Decile of deprivation; Key: A=Asian; B=Black; C=Chinese; O=Other/Mixed; R=Arab; W=White

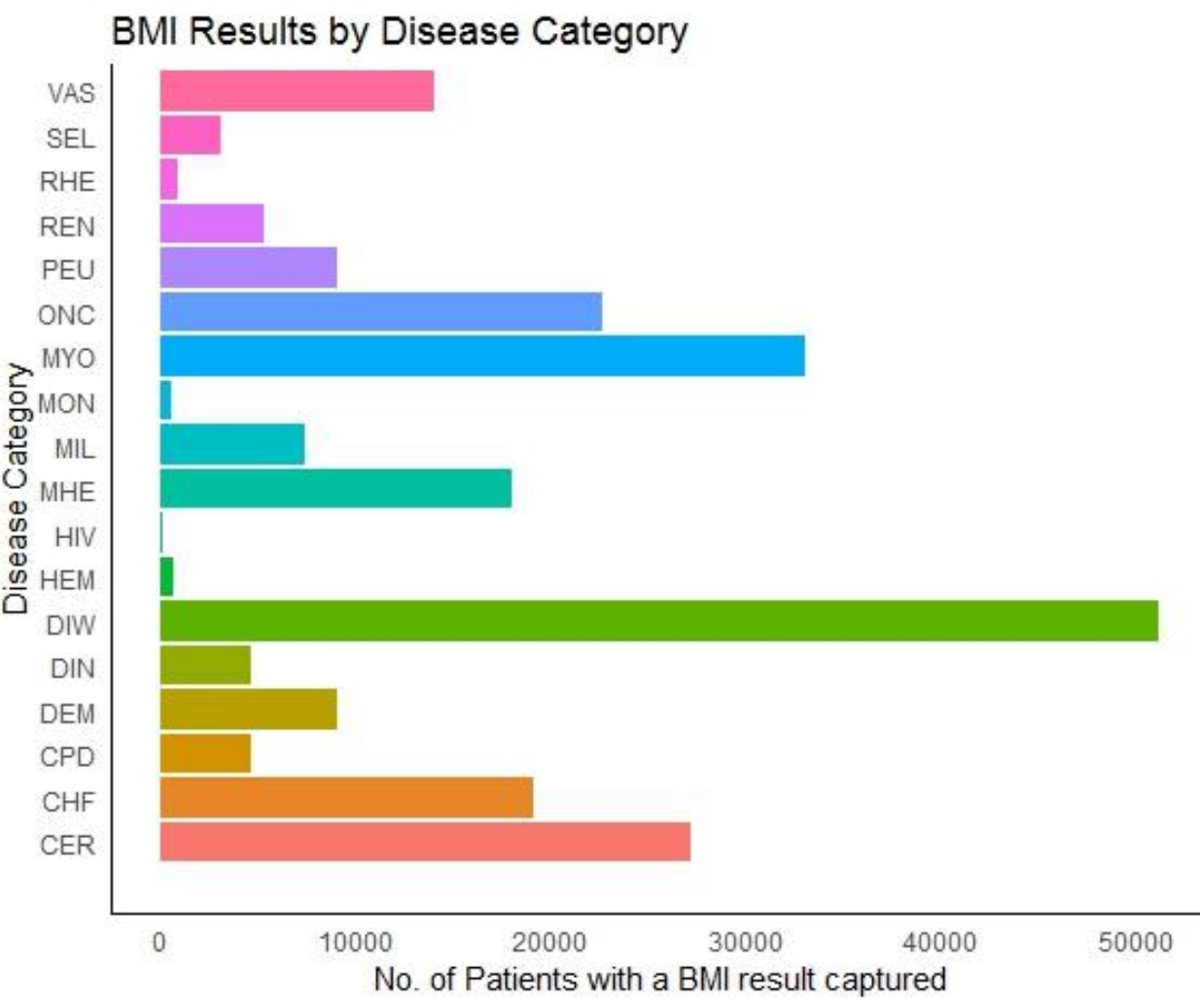


Figure 2: BMI capture by disease category
Key: VAS=Peripheral Vascular Disease; SEL=Moderate or Severe Liver Disease; RHE=Rheumatic Disease; REN=Renal Disease; PEU=Peptic Ulcer Disease; ONC=Any malignancy; MYO=Myocardial Infarction; MON=Metastatic solid tumour; MIL=Mild Liver Disease; MHE=Mental Health; HIV=AIDS/HIV; HEM=Hemiplegia or Paraplegia; DIW=Diabetes without chronic complication; DIN=Diabetes with chronic complication; DEM=Dementia; CPD=Chronic Pulmonary Disease; CHF=Congestive Heart Failure; CER=Cerebrovascular Disease

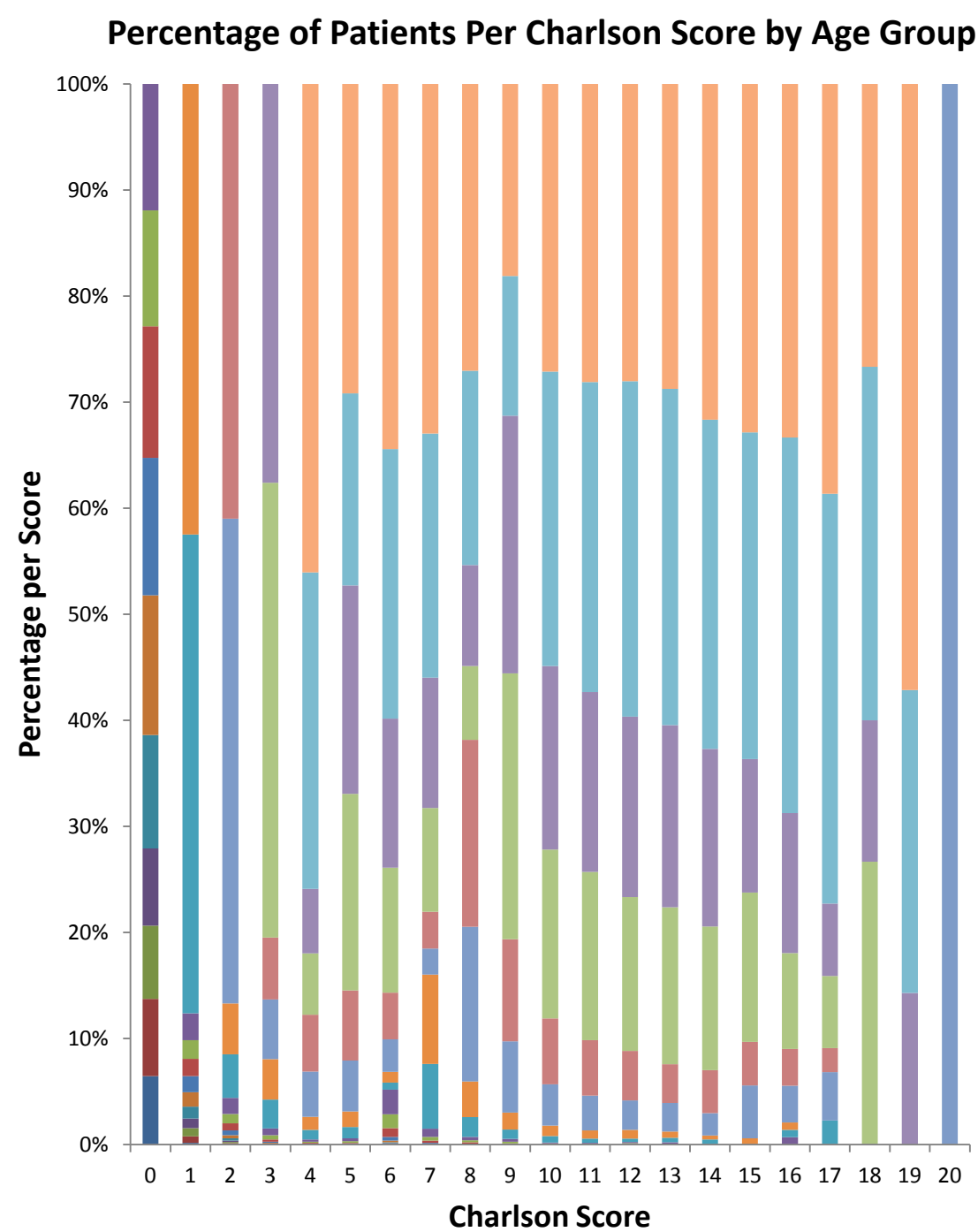


Figure 3: Percentage of living patients within each Age Group by Charlson Score

Dataset	No. Records	Included Individuals	No. Living	% of living GGC population
BMI	6,010,570	215,507	164,688	11.2
Ethnicity	1,122,199	884,065	761,913	51.8
Charlson Score	2,142,652	2,142,652	1,447,208	98.3
Smoking Status	288,851	220,371	186,403	12.7

Table 2: Patients found and included in the derived dataset as a percentage of the 1,471,812 living NHSGGC population

OUTCOMES/RESULTS

Derived datasets were created to hold consolidated data on BMI measurements, Ethnicity coding, Charlson comorbidity scores and Smoking status (including pack years). It resulted in NHSGGC Safe Haven going from having no reliable source of data for BMI, Ethnicity, Charlson Score or Smoking data to having relevant data for 11.2%, 51.8%, 98.3% and 12.7% of living patients, retrospectively (Table 2).

The BMI Matrix holds data on 215,507 individuals. The administrative SMR datasets were searched to categorise the patients by disease and visualise the disease relevance of the data (Figure 2). This shows the disease area with the highest capture is Diabetes.

The Ethnicity Matrix holds data on 1,122,199 patients. 997,126 of these are living patients however 113,061 had ambiguous data and a decision on their ethnicity was unable to be made. The remaining 884,065 enabled us to create a visualisation of the distribution and composition of our health board ethnicity, as it is able to be currently captured (Figure 1). Ethnicity data will now be added from other datasets to try and increase the coverage.

The Charlson Matrix has great coverage of 98.3% of the living population. Figure 3 and Figure 4 shows the number and distribution of patients by Charlson Score and age category. These demonstrate that the older sections of our population generally have the highest scores, which would be expected.

The Smoking Matrix produced an interesting dataset that can be used for many purposes. Here, we have shown the distribution of lung cancer and COPD in the population captured (Figure 5), plotted against pack years. This displays an increase in likelihood of developing lung cancer and COPD with an increase in pack years.

These datasets are now being used in five pilot longitudinal research projects with large data extracts where they are enriching phenotypic knowledge of the subjects involved. They have already been used in seven feasibility screens and have proved to be useful in identifying potential recruits.

Tier	Dataset ID	Name
1	7	SCI Score (Laboratory tests)
1	12	SMR00 (Outpatients - general)
1	13	SMR01 (Inpatients - general)
1	14	SMR02 (Inpatients - maternity)
1	15	SMR04 (Inpatients - mental health)
1	35	SCI Diabetes (Diabetes Registry)
1	72	SMR05 (Cancer Registry)
2	1	PsyCIS (Psychosis Clinical Information System)
2	2	Weight Management Service (WM service records)
2	3	Athena (Heart Failure service records)
2	4	Rheumatology (Rheumatology service records)
2	5	Parkinson's Disease (Specific Neurodegenerative conditions service records)
2	11	Clozapine Cohort (Local specific cohort dataset)
2	16	Stratified Medicine (Oncology pathology research dataset)
2	24	TrakCare A&E
2	25	SERPR (Renal service records)
2	26	Spirometry (Spirometry studies for GGC patients)

Table 1: Safe Haven resources used to derive datasets (NHSGGC Safe Haven, 2018)

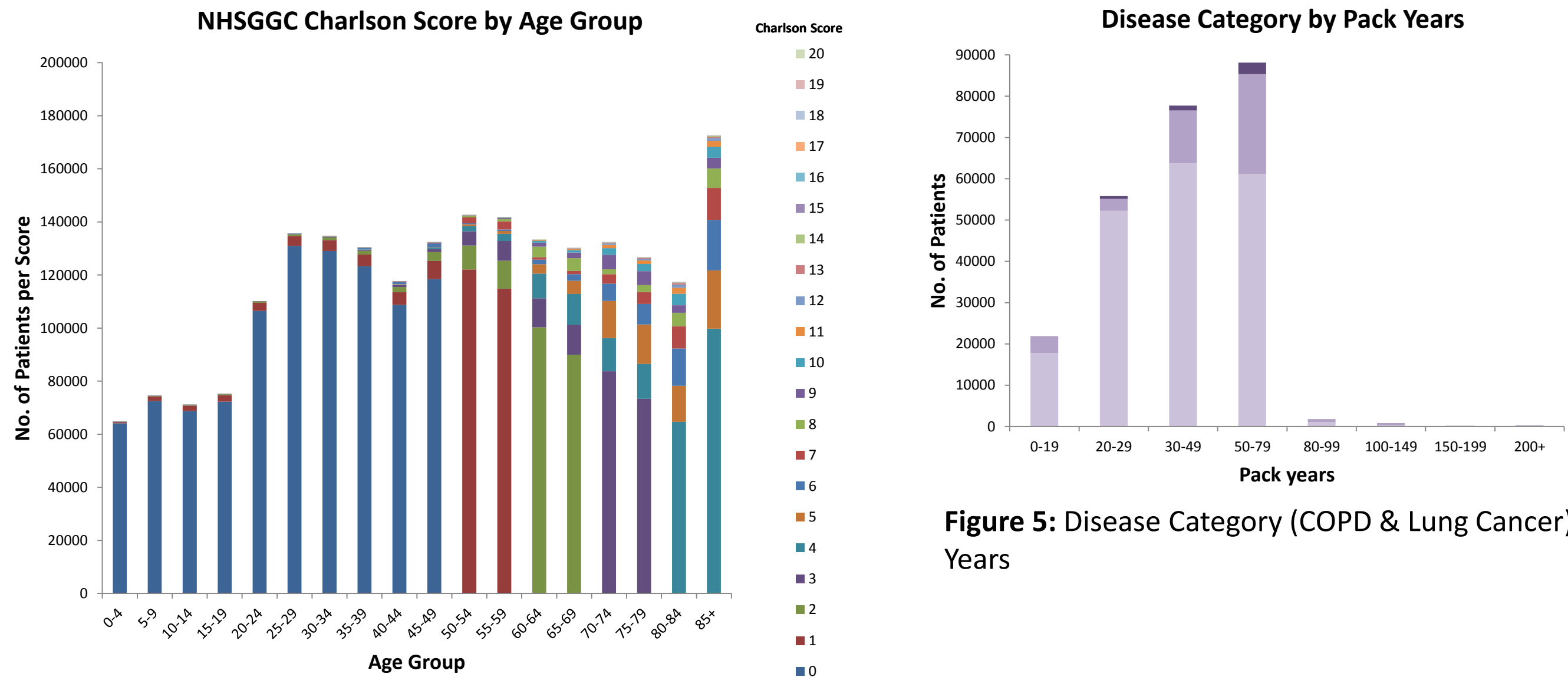


Figure 4: Number of living patients with each Charlson Score by Age Group

CONCLUSIONS

Local Safe Havens can add value to eResearch by creating derived datasets which help to build a better phenotypic and outcome picture of the subjects in a study. Multimorbidity is becoming increasingly important and therefore it is vital that as much data as possible is enabled to be captured and used. The practice of developing derived datasets should be encouraged where possible so Safe Havens can fortify research data projects and increase study recruitment.

REFERENCES

Deyo, R., Ciol, M.A., (1992), Cherkin, D.C. (1992), Adapting a Clinical Comorbidity Index for Use with ICD-9-CM Administrative Databases, *Journal of Clinical Epidemiology*, Vol 45, Issue 6, pp. 613-619 [Online] https://www.researchgate.net/publication/21547633_Adapting_a_Clinical_Comorbidity_Index_for_Use_With_ICD-9-CM_Administrative_Databases Accessed 04/05/18
NHSGGC Safe Haven, (2018), Available Datasets, viewed 22/10/18, <http://www.nhsggc.org.uk/about-us/professional-support-sites/nhsggc-safe-haven/your-research/available-datasets/>
Masters, N.J., (2018, September, 21), Smoking Pack years, viewed 04/05/18, <https://www.smokingpackyears.com/>
NICE (2018), Multimorbidity: clinical assessment and management, Accessed <https://www.nice.org.uk/guidance/ng56> 22/10/18
Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J., Saunders, D., Beck, C.A., Feasby, T.E., Ghali, W. (2005), Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data, *Medical Care*, Vol 43, Issue 11, pp.1130-1139 [Online] <https://www.ncbi.nlm.nih.gov/pubmed/16224307> Accessed 04/05/18
Wood, D.M., Mould, M.G., Ong, S.B.Y., Baker, E.H. (2005), “Pack year” smoking histories: what about patients who use loose tobacco?, *Tobacco Control*, Vol 14, pp. 141-142 [Online] <https://tobaccocontrol.bmj.com/content/14/2/141> Accessed 04/05/18

ACKNOWLEDGEMENTS

Dr John Haughney, Associate Director R&D, GP Director CRF, NHSGGC
Dr Nigel Masters, GP/Records Compliance Officer, Highfield Surgery, High Wycombe
Dr Deborah Morrison, Honorary Clinical Lecturer, General Practice & Primary Care, Institute of Health & Wellbeing, University of Glasgow
Dr David Morrison, Honorary Clinical Associate Professor, Institute of Health & Wellbeing, University of Glasgow